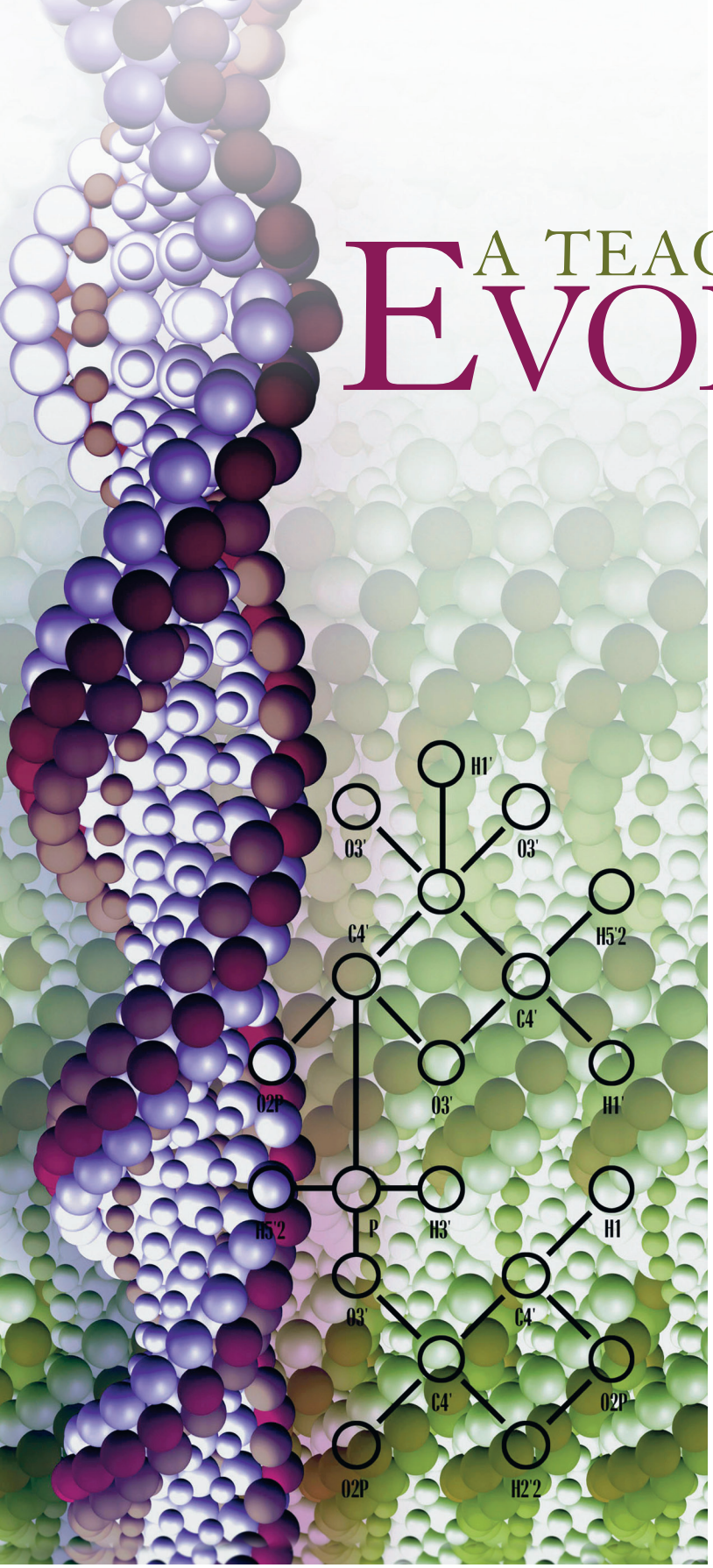
A TEACHING GUIDE TO EVOLUTION

Thomas G. Gregg, Gary R. Janssen, and J.K. Bhattacharjee



Discovering evolution through molecular evidence

Molecular evidence of evolutionary relationships

Universality of DNA structure

The molecular evidence in support of the evolutionary relationship among all living organisms is based upon clearly demonstrated, well understood, universally accepted scientific discoveries that can be verified again and again by students and teachers alike. Geneticists and

biochemists have learned that deoxyribonucleic acid (DNA) is the genetic material, and genes consist of nucleotide sequences in the double helix of DNA that specify the sequence of amino acids found in proteins. With the exception of some ribonucleic acid (RNA) viruses, DNA is the genetic material of all living organisms, including archaea, bacteria, protists, fungi, plants, and animals, including humans.

The primary structure of DNA is composed of deoxyribonucleotides linked to each other by phosphodiester linkage. The secondary structure of DNA consists of two antiparallel polynucleotide strands, which are hydrogen bonded to each other by the complementary base pairing between adenine (A) of one strand and thymine (T) of the other polynucleotide strand; likewise, a guanine (G) of one strand pairs with a cytosine (C) of the other polynucleotide strand. The general process of DNA synthesis, and the complex structure of DNA consisting of exactly the same four nucleotides, are conserved in all organisms and have not changed in more than three billion years of life on Earth.

Similarities among diverse organisms

Results from comparing the gene sequences for the small subunit ribosomal RNA, 16S in procaryotes or 18S in eucaryotes, from a wide variety of organisms shows that they all fall into three distinct clusters of similarity that define the three biological domains, Archaea, Bacteria, and Eucarya (Figure 1) (Woese 1998).

Despite this distinct grouping of all organisms into three domains they nonetheless share a number of features that indicate their common ancestry.

The complete sequencing of the human genome is now regarded as a major milestone in the history of scientific accomplishment (Lander et al. 2001; Venter et al. 2001). This feat and the technical advances that accompanied it have made it possible to sequence a number of other species as well and to make sequence comparisons among them. This accomplishment has opened a whole new field of scientific investigation called "comparative genomics." Comparisons of more than 60 species representing all three domains have revealed a high degree of sequence similarity and genetic relatedness among these widely diverse organisms.

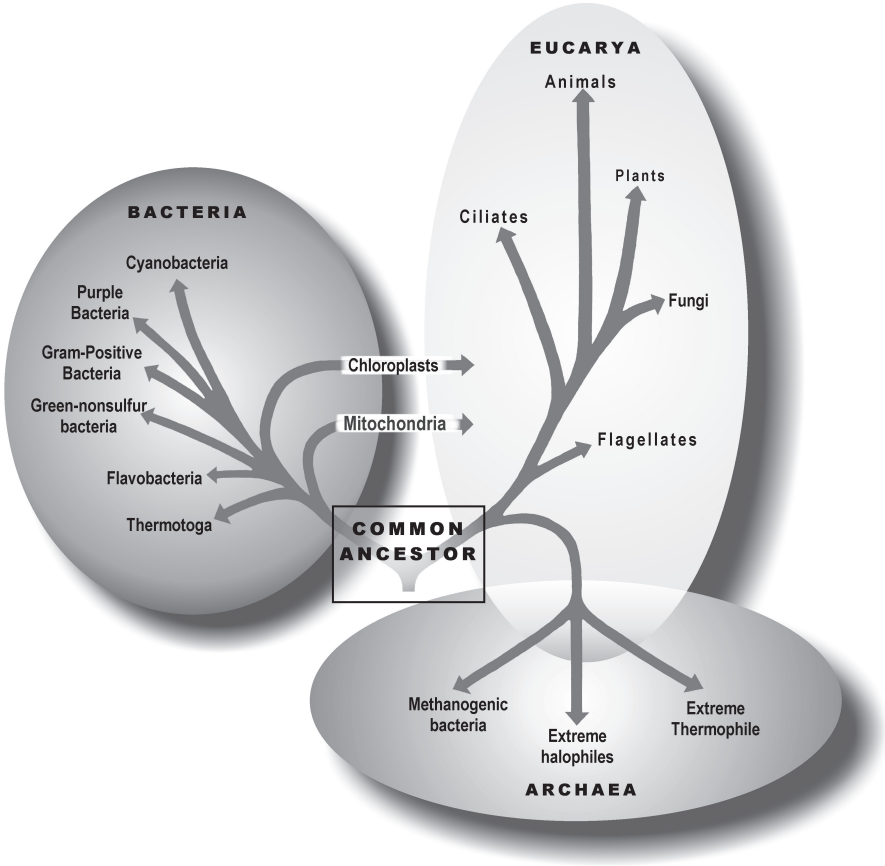
One result of particular interest is that eucaryotic nuclear genes are similar to those of the Archaea, whereas the mitochondrial genes of eucaryotes are similar to those of Bacteria. This is one line of evidence supporting the recent idea that the original relationship that gave rise to the first eucaryotes involved an organism in the Archaea engulfing a bacterium (Horiike et al. 2002; Martin and Muller 1998), although the still current paradigm is that mitochondria arose as endosymbionts in eucaryotes. Eukaryotic genes consist of numerous coding regions—exons—that are separated by noncoding regions—introns. The entire gene is transcribed into a pre-messenger RNA, from which the intron sequences are removed, joining the exons together into a single messenger RNA molecule.

Proteins consist of interrelated but somewhat independent functional domains. Research from the genome project shows that exons in genes correspond to functional domains in proteins. In human genes 90 percent of the

FIGURE 1

The three domains (Archaea, Bacteria, and Eucarya) that contain all forms of life found on Earth.

Phylogenetic relationships are based on homology of ribosomal RNA sequences. The distance between groups is proportional to their genetic differences and evolutionary relatedness (after Woese 1998, and McKane and Kandel 1996).



exons are homologous to exons found in *Drosophila* (fruit flies) and *Caenorhabditis* (nematode worm) (Rubin 2001). (Throughout this article homologous means sequences that are so similar that the similarities cannot be due to chance but are the result of common ancestry.) However, even though exons and protein domains are shared by widely diverse species they often are present in novel combinations and arrangements in different organisms.

For instance, another vertebrate that has been sequenced is the puffer fish. The puffer fish was chosen because it has the smallest vertebrate genome so far discovered and as such is considered to be most similar to the common ancestor of the vertebrates. Its genome is one-seventh the size of the human genome. However, puffer fish appear to have virtually all the exons that are present in humans. One reason for the difference in genome size is that the lineage leading to humans contains a great deal of duplication, both of genes and whole chromosome segments. Consequently, some of the puffer fish exons have been duplicated and reduplicated in humans and in many cases rearranged into new combinations. In this way it is possible for humans to have twice as many genes as puffer fish with the same number of exons.

Based on these observations from comparative genomics, vertebrate evolution has required the invention of very few new protein domains (Rubin 2001). Thus, one aspect of evolutionary change involves making new genes by rearranging functional domains into novel combinations. This process is called exon shuffling. From comparative genomics we have discovered that making new genes by exon shuffling is a very important source of genetic variation, beyond mutation and recombination, upon which natural selection can act. The mechanisms by which exon shuffling occurs are under active investigation (Long 2001).

At the level of whole genes, 60 percent of the human genes that encode proteins are homologous to genes from other organisms. Looking from the perspective of other organisms, 46 percent of yeast (*Saachromyces cervisiae*) genes, 43 percent of worm (*Caenorhabditis elegans*) genes, and 61 percent of fruit fly (*Drosophila melanogaster*) genes show sequence similarity with human genes (Lander et al. 2001). In summary, the high degree of conservation of both genes and exons among widely diverse organisms from all three phylogenetic domains is strong evidence for their common ancestry.

Nonfunctional sequences

Perhaps even stronger evidence for relatedness among diverse organisms than similarity among functional genes comes from similarity in DNA sequences that have no function. One category of nonfunctional sequence is the pseudogene. There are two kinds of pseudogenes. One kind arises from a gene duplication followed by mutations to stop codons in one of the duplicates with the other retaining the original function. The other kind of pseudogene is recognized because they, like messenger RNA, have a poly A tail sequence and lack a promoter sequence and introns. Without a promoter they cannot be transcribed. These are called processed pseudogenes because they evidently arise through reverse transcription of messenger RNA into double-stranded DNA, which then is incorporated into the genome.

To date, 2909 processed pseudogenes have been recognized in the human genome (Lander et al. 2001). The following pseudogene example supports relatedness and

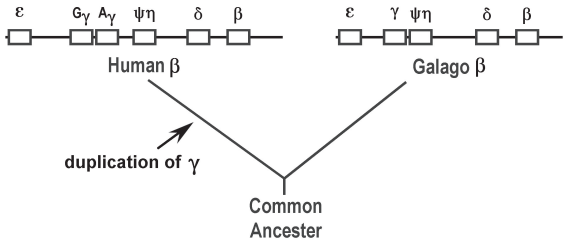
common ancestry. Figure 2A shows the beta hemoglobin gene cluster for two of the most distantly related primates—humans and *Galago* (Bush Baby). The cluster consists of beta (β), delta (δ), pseudogene eta (Ψη), gamma (γ), and epsilon (ε) genes. By sequence similarity, these genes are seen to have arisen by a series of duplications followed by mutational divergence, giving rise to their present functional differences that are expressed at different stages of embryonic, fetal, and adult development. The times at which the various hemoglobin duplications occurred are shown in Figure 2B. Of particular evolutionary interest is the conserved presence and conserved position of the functionless Ψη in all primate species (Goodman 1999). By far the most plausible explanation for this observation is that these species are all related and that the Ψη was present in the common ancestor.

If LINEs in different species are homologous (by sequence similarity) this is very strong evidence that the species share a common ancestor in which the LINE first became established. Otherwise the virus would have to have become defunct independently in each of the species,

FIGURE 2A

Beta hemoglobin gene cluster as it occurs in primates.

The pseudogene eta (ψη) shown here is also present in all other primates indicating their relatedness. (Note: Some textbook figures do not show the pseudogene.) Humans have acquired a duplication of the gamma gene, one designated with an A and the other G. These stand for adenine and guanine, the single nucleotide by which they differ. This duplication is present in all the simians (apes and monkeys) as well as in humans, but is not present in other primates. It is extremely unlikely that there is a physiological need for two copies of the gene in this group of primates but none of the others. Thus we conclude that monkeys, apes, and humans share this duplication because it arose in the common ancestor.

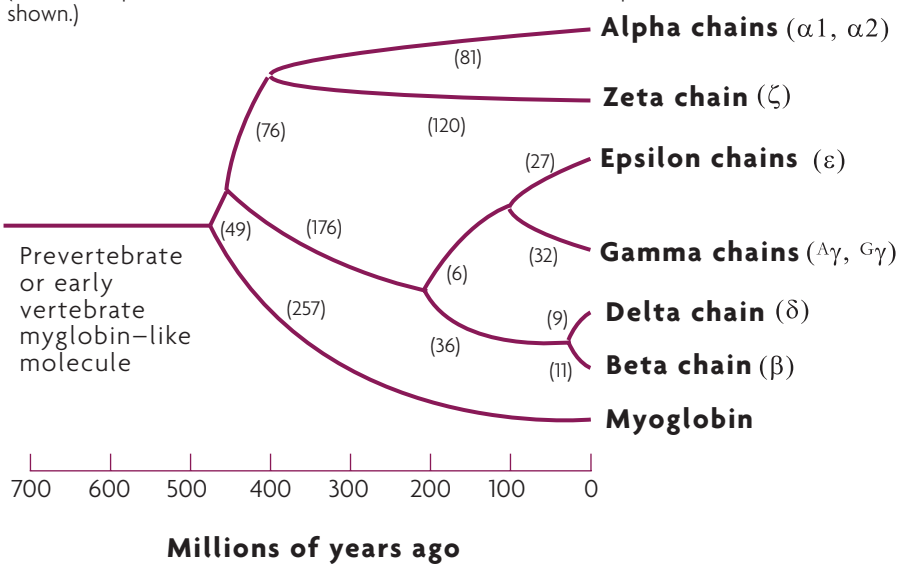


Another category of functionless DNA involves long interspersed nucleotide elements (LINEs). These are not entirely noncoding but they are nonfunctional as far as the individual organism is concerned. In the human genome there are three families of LINES: LINE 1, LINE 2, and LINE 3. There are 516 000 copies of LINE 1, 315 000 of LINE 2, and 37 000 of LINE 3 (Lander et al. 2001). Lines are now recognized to be retrotransposons, which are generally considered to be defunct retroviruses. Retroviruses are RNA viruses that—when they enter cells—cause their RNA to be converted into a double-stranded DNA molecule that is inserted into the host's DNA from which it controls the production of new viruses. LINEs retain the viral genes that allow their transcripts to be reverse transcribed into DNA and inserted into chromosomes, but have lost the genes for making coat proteins and escaping from the cell. In this way they transpose themselves into new chromosomal locations in a retrovirallike manner and so are called retrotransposons. This explains the thousands of copies.

FIGURE 2B

Evolutionary history of the duplications in the hemoglobin family.

(Numbers in parentheses are the estimated number of nucleotide replacements in the branches shown.)



Strickberger, *Evolution,* 2000: Jones and Bartlett Publishers, Sudbury, Mass. *www.jbpub.com.* Reprinted with permission.

which for even two species is considered to be very unlikely. The LINE 1 sequence is present not only in humans but in every species of mammal examined to date (Smit et al. 1995). Not only does this show indisputably that all mammals are related,

, but also that LINE 1 is very old.

In addition to LINEs in genomes of various organisms, there are short interspersed nucleotide elements (SINEs), which also can move around and are present in multiple copies. There are three families of SINEs in the human genome, collectively consisting of 1 558 000 copies (Lander et al. 2001).

Let us consider an example involving recent interest in the relationship of whales to other mammals. There are several specific LINEs and SINEs that are shared by whales and artiodactyls (the order of cloven-hoofed mammals), which are not present in other mammals. The likelihood that even one of these shared sequences arose independently in all of these species is virtually zero. Consequently, the conclusion is inescapable that all artiodactyls species are related to each other and also to whales (Nikaido, Rooney, and Okada 1999). That is, they share a common ancestor from whom they inherited the LINEs and SINEs that they share. Moreover, once a LINE or SINE is established in a specific location it does not leave. Thus if a particular LINE or SINE is found in identical locations in each of two species that also is an indication of common ancestry. One SINE, Chr-1, is found in exactly the same position, to the nucleotide pair, in four different genes of whales and hippopotami. Chr-1 is not found in any of these locations in any other artiodactyl species. Thus whales and hippopotami clearly are each other's closest living relatives (Nikaido, Rooney, and Okada 1999), necessitating the establishment of a new suborder, Cetartiodactyla.

The presence of conserved, but nonfunctional, sequences in a variety of different organisms is more consistent with an ongoing evolutionary process

Universality of triplet code and translation process

Protein synthesis is a highly conserved fundamental life process and is responsible for most metabolic activities and phenotypes in all organisms. All living organisms synthesize proteins; it is a universal characteristic of life. Moreover, all organisms do it the same way, using messenger RNA, transfer RNA, ribosomes, and the same 20 amino acids. The triplet genetic code is virtually universal. The few exceptions, such as in a few ciliates and the mitochondria of some organisms, are all easily understood as single step changes from the standard code. Also, the processes of transcription and translation for protein synthesis are highly conserved among all forms of life, indicating that very little change has taken place since the time of the common ancestor over three billion years ago. Although the molecules involved in transcription and translation have undergone some changes in nucleotide and amino acid sequence, their function has not changed.

Differences in ribosomal RNA sequences determined the three domains, while they continue to carry out the same function in the small ribosomal subunit. The universality of both the genetic code and the method of protein synthesis is another strong argument for the relatedness of all organisms because such a high degree of complex metabolic similarity did not evolve millions of times independently. Similarity and conservation of complex processes exist not just at the metabolic level, but also at every level from the molecular to the morphological. What is the explanation for complex similarities? The argument for evolution, based on genetic descent with modification, is powerful and persuasive.

Biochemical evidence of evolutionary relationships

Metabolic pathways

Biochemists and geneticists have accumulated experimental evidence since the 1930s that all organisms contain interdependent biochemical and metabolic pathways that fail to work if one component becomes defective.

Evolutionists, argue that complex pathways have been built up by adding one step at a time. One line of evidence for stepwise addition is the fact that genes for the enzymes in a pathway are frequently so similar in sequence that they clearly have arisen as a series of gene duplications from one original gene, and this would necessarily have occurred in stepwise fashion (Chothia et al. 2003).

The basic chemical composition of cells (i.e., carbon, hydrogen, oxygen, nitrogen, phosphorus, and sulfur) and many aspects of cellular metabolism are also highly conserved. For instance, virtually all present-day organisms, both aerobic and anaerobic, carry out glycolysis, the conversion of 6-carbon glucose molecules to two 3-carbon pyruvic acid molecules. This is because the glycolytic metabolism of glucose to pyruvic acid has been conserved in all three domains from the time of the common ancestor.

The Krebs Cycle and its associated electron transport system are present in all aerobic eucaryotes because the reactions are carried out by mitochondria. Since virtually all biologists agree that mitochondria arose through bacterial endosymbiosis, this event must have occurred in the common ancestor of all present day eucaryotes.

At the metabolic level, just as at the levels of morphology, anatomy, and development, the same complex traits require similar sets of genes to produce them. Genes are passed from one generation to the next, therefore the most plausible explanation for complex genetic similarities is common ancestry.

Genetic variability of proteins

Genes and their encoded proteins that perform the same metabolic functions in different organisms are similar but in most cases not identical with respect to their DNA and protein sequences. For example, the cytochrome C protein that performs the same electron transport function in horse and cow mitochondria is very similar, but not identical. In fact, comparing cytochrome C among 60 different species examined revealed that only 27 amino acid residues are identical in all 60 while more than 60 residues differ among them (the latter number is not exact because there are slight differences in the length of the molecule in some species).

However, the degree of similarity among amino acid sequences in cytochrome C corresponds closely to the phylogenetic relationships based on other criteria. That is, mammalian sequences are more similar to each other than to any reptilian sequence and vice versa, and so on. Similar patterns exist for other proteins that have been compared among species. For instance, the human hemoglobin alpha chain differs from that of sharks by 79 out of 141 amino acids, from a bony fish (carp) by 68, an amphibian (newt) by 62, a bird (chicken) by 35, a horse by 18, and a chimpanzee by 0 (Figure 3).

Mammals are more similar to each other than any mammal is to a bird.

But evolution explains them perfectly. Such genetic and biochemical changes have taken place over millions of years of evolutionary time by a combination of selective improvements and fixation of neutral mutations that have no effect on protein function.

On the other hand, despite the enormous potential for flexibility, comparison of the amino acid sequences of nine different proteins in humans and chimpanzees (including hemoglobin and cytochrome C) reveals a total of only five amino acid differences (Strickberger 2000). This is very strong evidence that the two species are very closely related and are each other's closest relative.

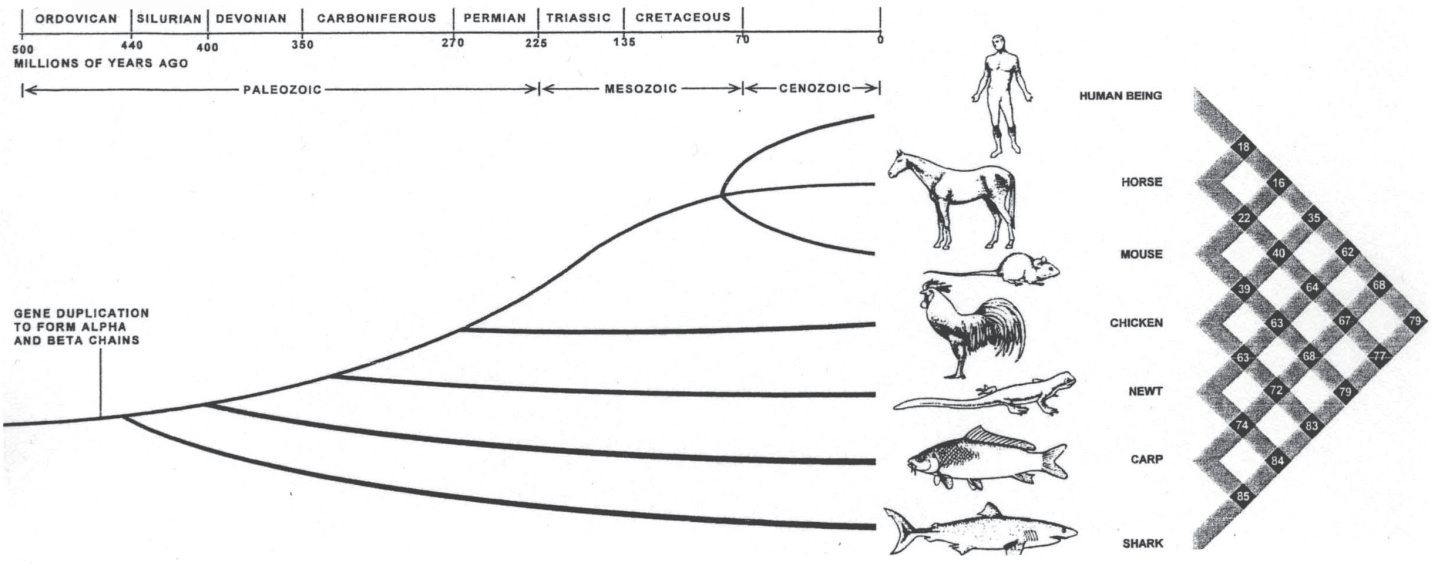
Genetics of evolution

With Mendelism it became clear that genes determine the characteristics of individual organisms and that they occasionally mutate. This much knowledge was enough to illuminate the relationship between mutation and natural selection; between genetics and evolution. Indeed, Mendel's laws of genetic inheritance, rather than some fossil intermediate between two presently existing species, were Darwin's missing link. The understanding

FIGURE 3

The number of amino acid differences in the alpha hemoglobin chain between the representative species.

The number of amino acid differences between any two species is read at the intersection of the diagonals. For instance, shark differs from carp by 85, from newt by 84, and from chicken by 83 (from Kimura 1979).



that genetic mutations produce the genetic variation in natural populations upon which natural selection acts was consolidated in the 1930s (the neo-Darwinian synthesis). Mutation and natural selection were finally seen as complementary factors, rather than competing mechanisms in the origin of species. We could see how, in the words of the influential evolutionary biologist Theodosius Dobzhansky (1951), "evolution is a change in the genetic composition of populations."

Out of the neo-Darwinian synthesis grew the field of population genetics. Geneticists strove to determine how much genetic variation existed in natural populations and what factors other than mutation and natural selection, such as migration, geographic isolation, and chance (genetic drift), might interact with natural selection to change gene frequencies. Mathematical formulae were developed to calculate the effects of these various factors on gene frequencies in populations, and various modes of speciation were investigated. These studies continue today and have been joined by ecologists who study specific interactions between organisms and their environments to determine environmental effects on gene frequencies. All of this began before we learned that genes specify proteins, or that genes were DNA molecules.

Gene duplication and evolution

Another question that arose was the origin of new genes. Obviously more complex organisms need more genetic information than simple organisms, although we were surprised to learn that humans have only slightly more than twice as many genes as *Drosophila* (Venter et al. 2001). If complex organisms evolved from simple ones where did the new genes come from? It was hypothesized from the outset that on rare occasions a small section of chromosome, along with the genes it contains, is duplicated. The mechanism or mechanisms for this are not entirely clear, perhaps by unequal crossing over, but their existence is indisputable. After duplication, one copy of a gene can maintain the original function while the other is free to mutate to a different function.

There was abundant evidence from *Drosophila* for small duplications, and as shown in Figure 2A (p. 27), the occurrence of small duplications can be seen directly in the globin gene families of vertebrates. Moreover if we compare the relative differences among the various globin genes, as was done for humans, it is possible to establish the evolutionary history of the globin family. From these differences we can deduce that the alpha chain gene arose from a duplication of the myoglobin gene and the beta chain gene from a duplication of the alpha chain gene, and approximately how long ago the duplications occurred (Figure 2B, p. 27). Additional duplications then occurred in the alpha and beta gene clusters. The alpha cluster now consists of alpha 1, alpha 2, and zeta, and the beta cluster consists of beta, delta, gamma, and epsilon. If a group of animals, such as the primates in Figure 2A (p. 27), share a complex gene cluster that arose by a series of duplications it provides overwhelming evidence that the cluster existed in a common ancestor and that the species in the group are biologically related

Exon shuffling and evolution

We have already indicated that exon shuffling as a method for making new genes is a reality. But let us consider a specific example that illustrates both exon shuffling and serial duplication. It involves the genes encoding the enzymes for the mammalian bloodclotting cascade,

By sequence analysis it has been shown that these genes contain exons homologous to exons in the trypsin gene and others homologous to those in the gene for epidermal growth factor (Doolittle 1993). Each of the six enzymes in the cascade is, like trypsin, produced in an inactive form that is activated by having a segment of its amino acids removed. Each active enzyme in the cascade is, again like trypsin, a serine protease that functions to activate the next enzyme in the cascade.

Finally, prothrombin is converted to thrombin, which converts fibrinogen into fibrin and the clot forms. For the proenzymes to be cleaved they must be bound to cell membranes in the damaged tissue. This takes place because there is a membrane receptor called "tissue factor" that binds to the epidermal growth factor end of the proenzymes. And just as the genes for the enzymes in the cascade are, by sequence similarity, homologous to genes with other but related functions, the gene for tissue factor is homologous to the gene for epidermal growth factor receptor. The gene for fibrinogen is homologous to a gene for actin, a widespread contractile molecule, best known for its presence in muscles. From these observations alone the bloodclotting cascade looks more like a case of evolutionary tinkering, working with what is already available, When one takes into consideration the high degree of sequence similarity between the enzymes in the cascade it seems obvious that the first bloodclotting gene arose by exon shuffling and subsequent genes of the cascade arose from simple gene duplication followed in each case by mutation to new, but related, functions.

Genome duplication and evolution

In order for exon shuffling to work there must be duplicate sets of genes if the original functions are to be retained. The human genome project has shown that the human genome is extensively duplicated, so extensively that it is debated whether the duplications arose from successive episodes of polyploidy (i.e., the doubling of whole sets of chromosomes) or by some other mechanism.

Gene duplication and selection for novel function likely has occurred also within bacteria. For example, sequence data reveals that the *Escherichia coli (E. coli)* genome is 4.6 million base pairs, the *Pseudomonas aeruginosa* genome is 6.2 million base pairs, and the *Streptomyces coelicolor* genome is 8.6 million base pairs. While all of these organisms are procaryotic, the range in genome size suggests that habitat diversity is selected for larger genomes with increased functional diversity. Also, the *E. coli* genome evolved to be 90 percent coding sequence whereas more than 90 percent of the human genome is noncoding. An additional major difference is that the coding region of a procaryotic gene is a continuous exon but the coding region of a higher eukaryotic gene is segmented with introns and exons. The promoter region is also very different between procaryotic and eucaryotic genes, as demanded by the greater regulatory sophistication required by the metabolic and developmental complexity of eucaryotes.

Thomas G. Gregg (e-mail:* greggtg@muohio.edu*) is an associate professor in the Department of Zoology, and Gary R. Janssen (e-mail:* janssegr@muohio.edu*) and J.K. Bhattacharjee (e-mail:* bhattajk@muohio.edu*) are both professors in the Department of Microbiology, all at Miami University, Oxford, OH 45056.

References

Asfaw, B., W.H. Gilbert, Y. Beyene, W.K. Hart, P.R. Renine, G. Woldegabriel, E.S. Vrba, and T.D. White. 2002. Remains of Homo erectus from Bouri, Middle Awash, Ethiopia. *Nature* 416:317–320.

Behe, M. 1996. *Darwin's Black Box: The Biochemical Challenge to Evolution*. New York: The Free Press.

Chothia, C., J. Gough, C. Vogel, and S.A. Techmann. 2003. Evolution of the protein repertoire. *Science* 300:1701–1703.

Dobzhansky, T. 1951. *Genetics and the Origin of Species*. New York: Columbia University Press.

Doolittle, R.E. 1993. The evolution of vertebrate blood coagulation: A case of yin and yang. *Thrombosis and Haemostasis* 70:24–28.

Goodman, M. 1999. The genomic record of humankind's evolutionary roots. *American Journal of Human Genetics* 64:31–39.

Hardison, R., Hardison, R., J.L. Slighton, D.L. Gumucio, M. Goodman, N. Stojanovic, and W. Miller. 1997. Locus control regions of mammalian beta-globin gene clusters: Combining phylogenetic analyses and experimental results to gain functional insights. *Gene* 205:73–94.

Horiike, T., K. Hamada, S. Kayana, and T. Shinozawa. 2001. Origin of eukaryotic cell nuclei by symbiosis of Archea in Bacteria is revealed by homology hit analysis. *Nature Cell Biology* 3:210–214.

Kimura, M. 1979. The neutral theory of molecular evolution. *Scientific American* 241(5): 94–104.

Lander, E.S., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921.

Long, M. 2001. Evolution of novel genes. *Current Opinion in Genetics and Development* 11:673–680.

Martin, W., and M. Muller. 1998. The hydrogen hypothesis for the first eucaryote. *Nature* 392:37–41.

McKane, L., and J. Kandel. 1996. *Microbiology*. New York: McGraw-Hill.

Nikaido, M., A.P. Rooney, and N. Okada. 1999. Phylogenetic relationships among cetartiodactyls based on short and long interspersed elements: Hippopotamuses are the closest extant relatives of whales. *Proceedings of the National Academy of Sciences* 96(18): 10261–10266.

Rubin, G.M. 2001. Comparing species. *Nature* 409:820–821.

Smit, A., C. Toth, A. Riggs, and J. Jurga. 1995. Ancestral mammalian-wide subfamilies LINE 1 repetitive sequences. *Journal of Molecular Biology* 246:401–417.

Strickberger, M.W. 2000. *Evolution*. 3rd ed. Sudbury, Mass.: Jones and Bartlett.

Venter, C., et al. 2001. The sequence of the human genome. *Science* 291:1304–1351.

Woese, C.R. 1998. The universal ancestor. *Proceedings of the National Academy of Sciences* 95:6854–6859.

Acknowledgment

The authors thank Richard Edelmann of Miami University for excellent assistance with the figures.

On the Web

For a glossary of terms used in this article, visit *www.nsta.org/highschool*.

NSTA Connection

This fall, look for *Evolution in Perspective: The Science Teacher's Compendium*, published by NSTA Press. Visit *www.nsta.org/pubs/nstapress* for this and other NSTA books on evolution including *The Creation Controversy & The Science Classroom*.